

When does Active Learning Work?

Lewis P. G. Evans¹, Niall M. Adams^{1,2}, and Christoforos Anagnostopoulos¹

¹ Department of Mathematics, Imperial College London

² Heilbronn Institute for Mathematical Research, University of Bristol

Abstract. Active Learning (AL) methods seek to improve classifier performance when labels are expensive or scarce. We consider two central questions: Where does AL work? How much does it help? To address these questions, a comprehensive experimental simulation study of Active Learning is presented. We consider a variety of tasks, classifiers and other AL factors, to present a broad exploration of AL performance in various settings. A precise way to quantify performance is needed in order to know when AL works. Thus we also present a detailed methodology for tackling the complexities of assessing AL performance in the context of this experimental study.

Keywords: classification, active learning, experimental evaluation, algorithms

1 Introduction

Active Learning (AL) is an important sub-field of classification, where a learning system can intelligently select unlabelled examples for labelling, to improve classifier performance. The need for AL is often motivated by practical concerns: labelled data is often scarce or expensive compared to unlabelled data [9].

We consider two central questions: Where does AL work? How much does it help? These questions are as yet unresolved, and answers would enable researchers to tackle the subsequent questions of how and why AL works.

Several studies have shown that it is surprisingly difficult for AL to outperform the simple benchmark of random selection ([3,8]). Further, both AL methods and random selection often show high variability which makes comparisons difficult. There are many studies showing positive results, for example [9,5]. Notably there are several studies showing negative results, for example [2,8]. While valuable, such studies do not permit any overview of where and how much AL works. Moreover, this contradiction suggests there are still things to understand, which is the objective of this paper.

We take the view that a broader study should try to understand which factors might be expected to affect AL performance. Such factors include the classification task and the classifier; see Section 2.3. We present a comprehensive simulation study of AL, where many AL factors are systematically varied and subsequently subjected to statistical analysis.

Careful reasoning about the design of AL experiments raises a number of important methodological issues with the evaluation of AL performance. This

paper contributes an assessment methodology in the context of simulation studies to address those issues.

For practical applications of AL, there is usually no holdout test dataset with which to assess performance. That creates major unresolved difficulties, for example the inability to assess AL method performance, as discussed in [8]. Hence this study focusses on simulated data, so that AL performance can be assessed.

The structure of this paper is as follows: we present background on classification and AL in Section 2, then describe the experimental method and assessment methodology in Sections 3 and 3.1. Finally we present results in Section 4 and conclude in Section 5.

2 Background

This section presents the more detailed background on classification and AL.

2.1 Classification

Notationally, each classification example has features \mathbf{x}_i and a corresponding label y_i . Thus each example is denoted by $\{\mathbf{x}_i, y_i\}$, where \mathbf{x}_i is a p -dimensional feature vector, with a class label $y_i \in \{C_1, C_2, \dots, C_k\}$.

A dataset consists of n examples, and is denoted $D = \{\mathbf{x}_i, y_i\}_1^n$. A classifier is an algorithm that predicts classes for unseen examples, with the objective of good generalisation on some performance measure. A good overview of classification is provided by [7, Chapter 1,2].

2.2 Active Learning

The context for AL is where labelled examples are scarce or expensive. For example in medical image diagnosis, it takes doctors' valuable time to label images with their correct diagnoses; but unlabelled examples are plentiful and cheap. Given the high cost of obtaining a label, systematic selection of unlabelled examples for labelling might improve performance. An AL method can guide selection of the unlabelled data, to choose the most useful or informative examples for labelling. In that way the AL method can choose unlabelled data to best improve the generalisation objective. A small set of unlabelled examples is first chosen, then presented to an expert (*oracle*) for labelling.

Here we focus on batch AL; for variations, see [9,4]. A typical scenario would be a small number of initially labelled examples, a large pool of unlabelled examples, and a small budget of label requests. An AL method spends the budget by choosing a small number of unlabelled examples, to receive labels from an oracle.

An example AL method is uncertainty sampling using Shannon Entropy (denoted SE). SE takes the entropy of the whole posterior probability vector for

all classes. Informally, SE expresses a distance metric of unlabelled points from the classifier decision boundary.

$$Entropy = - \sum_{i=1}^k p(y_i|\mathbf{x}_j) \times \log(p(y_i|\mathbf{x}_j)).$$

Another example AL method is Query By Committee (denoted QBC), described in section 4.2.

The oracle then satisfies those label requests, by providing the labels for that set of unlabelled examples. The newly-labelled data is then combined with the initially-labelled data, to give a larger training dataset, to train an improved classifier.

The framework for AL described above is batch pool-based sampling; for variations see [4,9].

2.3 Active Learning Factors

Intuitively there are several factors that might have an important effect on AL performance. An experimental study can vary the values of those factors systematically to analyse their impact on AL performance.

One example of an AL factor is the nature of the classification task, including its difficulty and the complexity of the decision boundary. The classifier can be expected to make a major difference, for example whether it can express linear and non-linear decision boundaries, and whether it is parametric. The smoothness of the classification task input, for example continuous or discretised, might prove important since that smoothness affects the diversity of unlabelled examples in the pool. Intuitively we might expect a discretised task to be harder than a continuous one, since that diversity of pool examples would decrease. Other relevant factors include the number of initial labels ($N_{initial}$) and the size of the label budget (N_{budget}).

Some of these factors may be expected to materially determine AL performance. How the factors affect AL performance is an open question. This experimental study evaluates AL methods for different combinations of factor values, i.e. at many points in factor space. The goal here is to unravel how the factors affect AL performance. A statistical analysis of the simulation study reveals some answers to that question, see Section 4.1.

Below the factor values are described in detail.

Four different simulated classification tasks are used, to vary the nature and complexity of the classification problem. We restrict attention to binary classification problems. Figure 1 shows the classification tasks. These tasks are created from mixtures of Gaussian clusters. The clusters are placed to create decision boundaries, some of which are simple curves and others are more involved. In this way the complexity of the classification problem is varied across the tasks.

Still focussing on the classification task, task difficulty is varied via the Bayes Error Rate (BER). Input smoothness is also varied, having the values continuous, discretised, or a mixture of both. BER is varied by modifying the Gaussian

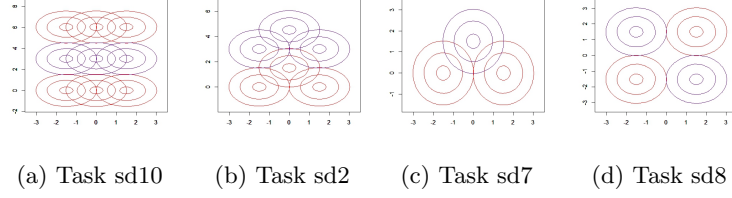


Fig. 1: Density contour plots to elucidate the classification problems

clusters for the problems; input smoothness is varied by transforming the realised datasets.

Another factor to vary is the input dimension p , by optionally adding extra dimensions independent of the class. An interaction is expected between p and the initial amount of labelled data $N_{initial}$, since higher dimensional data should require more datapoints to classify successfully.

Four different classifiers were used: Logistic Regression (LogReg), Quadratic Discriminant Analysis (QDA), Random Forest (RF) and Support Vector Machines (SVM), to provide a variety of classifiers: linear and non-linear, parametric and non-parametric. These classifiers are described in [7]. The default parameters for RF are the defaults from R package RandomForest version 4.6-7; the default parameters for SVM are the defaults from R package RWeka version 0.4-14 (the complexity parameter C is chosen by cross-validation, the kernel is polynomial).

The amount of initial labelled data $N_{initial}$ is also varied. This factor is expected to be important, since too little data would give an AL method nothing to work with, and too much would often mean no possible scope for improvement.

The AL factors are summarised in Table 1.

Table 1: Active Learning Factors

Name	Values
Classification Task	sd10, sd2, sd7, sd8 (see Figure 1)
Task Input Type	Continuous, Discretised, Mixed
Task Input Dimension	2, 10
Classifier	LogReg, QDA, RF, SVM
$N_{initial}$	10, 25, 50, 100
Bayes Error Rate	0.1, 0.2, 0.35
Classifier Optimum Error Rate	[inferred]
Space for AL	[inferred]

The optimum error rate for the classifier on a specific task is evaluated experimentally, by averaging the results of several large train-test datasets, to provide a ceiling benchmark.

We also consider the potential space for AL to provide a performance gain. In the context of simulated data all labels are known, and some labels are hidden to perform the AL experimental study. The classifier that sees all the labelled data provides a ceiling benchmark, the score S_{all} . The classifier that sees only the initially labelled data provides a floor benchmark, the score $S_{initial}$. To quantify the scope for AL to improve performance, we define the space for AL as a ratio of performance scores: $(S_{all} - S_{initial})/S_{all}$. This provides a normalised metric of the potential for AL to improve performance.

A Monte Carlo experiment varying these factors provides the opportunity to statistically analyse the behaviour of AL. To get to this point, both a careful experiment and a refined methodology of performance assessment are required.

3 Experimental Method

AL is applied iteratively in these experiments: the amount of labelled data grows progressively, as the AL method spends a budget chunk at each time point. We may choose to spend our overall budget all in one go, or iteratively, in smaller batches.

In that sense the experimental setup resembles that of the AL challenge described in [5]. We use this iteration over budget because it is realistic for practical AL applications, and because it explores the behaviour of AL as the number of labelled examples grows. Experiments consider AL methods SE and QBC.

To motivate our experimental method, we present the summary plots of the relative performances of AL and RS over time, see Figures 2a and 2b.

The experimental setup is as follows. Firstly, sample a pair of datasets $[D_{train}, D_{test}]$ from the classification task. To simulate label scarcity, split the training dataset into initially labelled data $D_{initial}$ and an unlabelled pool D_{pool} .

The output of one experiment can be described in a single plot, for example Figure 2a. That figure shows the trajectory of performance scores obtained from progressive labelling, as follows. At each time point the AL method chooses a small set of examples for labelling, which is added to the existing dataset of labelled data. This selection happens repeatedly, creating a trajectory of selected datasets from the unlabelled pool. Each time point gives a performance score, for example error rate, though the framework extends to any performance metric. This gives the overall result of a trajectory of scores over time, denoted \mathbf{S}_i : an empirical learning curve. Here i denotes the time point as we iterately increase the amount of labelled data, with $i \in [0, 100]$.

Given several instances of RS, we form an empirical boxplot, called a sampling interval. Figure 2a shows the trajectory of scores for the AL method, and the vertical boxplots show the sampling intervals for the scores for RS.

Once this iterative process is done, we obtain a set of scores over the whole budget range, denoted \mathbf{S}_i . Those scores are used to calculate various performance comparisons, specifically to see whether AL outperformed RS, see Section 3.1.

The AL method now has a score trajectory \mathbf{S}_i : a set of scores over the whole budget range. All trajectories begin at the floor benchmark score $S_{initial}$ and terminate at the ceiling benchmark score S_{all} . From the score trajectory \mathbf{S}_i a set of score differences $\delta\mathbf{S}_i$ is calculated via $\delta\mathbf{S}_i = \mathbf{S}_i - \mathbf{S}_{i-1}$. The need for and usage of the score differences is detailed in Section 3.1. The chosen AL method is evaluated alongside several instances of RS, the latter providing a benchmark. Experiments are repeated to generate several instances of RS, since RS shows substantial variability.

To illustrate the trajectories of the performance scores \mathbf{S}_i , Figure 2a shows those scores for the AL method SE and comparison with RS.

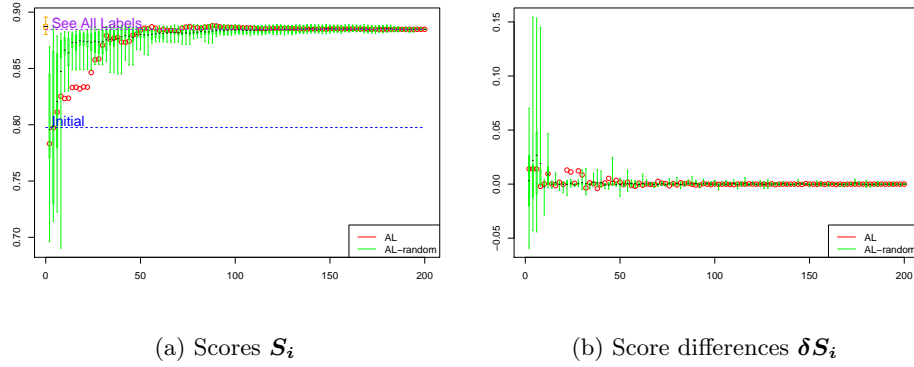


Fig. 2: Scores \mathbf{S}_i in subgraph (a) and score differences $\delta\mathbf{S}_i$ in subgraph (b), for AL method Shannon Entropy vs Random Selection

3.1 Methodology to evaluate AL performance

This section elucidates the difficulties with existing AL performance metrics, and contributes a novel assessment methodology to address those complexities. The primary goal of every AL performance metric is to quantify the AL performance gain from a given experiment, as a single scalar summary.

Any AL performance methodology must first address two preliminary issues: the benchmark for AL to outperform, and how to handle the variability of that benchmark. The first issue is to decide which benchmark should be used to compare AL methods against. One option is to compare AL performance to the initial classifier. However, that ignores the fact that the labelled dataset is larger in the case of AL: even random selection of further examples for labelling would be expected to improve performance on average, since the classifier sees a larger training dataset. Thus a better benchmark for AL is random selection (RS), that sees exactly the same amount of labelled data as the AL method.

The second issue concerns the high variability of the benchmark, given that experiments show RS to have high variability. The approach used here is to evaluate multiple instances of RS, to get a reasonable estimate of both location and dispersion of performance score. From those multiple instances we can form a sampling interval of the RS score, and thus capture its high variability.

Score trajectories under experimental budget iteration Having established the benchmark of RS, we consider the score trajectories in the experimental context of budget iteration, to better understand how to compare AL against its benchmark.

We begin with the score trajectories \mathbf{S}_i derived from the budget iteration process. The budget is iterated over the entire pool in 100 steps; during that iteration, the amount of labelled data grows from its minimum $N_{initial}$ to its maximum N_{train} . At each budget iteration step, the available budget is small compared to the total size of the pool. This is illustrated in Figure 2a.

Each score trajectory \mathbf{S}_i has significant autocorrelation, since each value depends largely on the previous one. To see this for the score trajectory, recall that for each budget iteration step, the available budget is small. Hence the score at one step S_i is very close to the score at the previous step S_{i-1} . Thus the scores \mathbf{S}_i only change incrementally with each budget iteration step, giving rise to high degrees of autocorrelation.

In contrast, the score differences $\delta\mathbf{S}_i$ are expected to be substantially less autocorrelated. This belief is confirmed experimentally by ACF graphs, which show significant autocorrelation for the scores but not for the score differences. This contrast matters when comparing different AL performance metrics.

Comparing AL performance metrics We now address different AL performance metrics, each designed to measure the performance of AL methods. Two common AL performance metrics are direct comparisons of the score trajectories, and the Area Under the Active learning curve (AUA) (see [5]).

The autocorrelation of score trajectories \mathbf{S}_i means that directly comparing two score trajectories is potentially misleading. For example, if an AL method does well against RS only for a small time at the start, and then does equally well, this would lead to the AL method’s score trajectory dominating that of the RS over the whole budget range. This would present a false picture of where the AL method is outperforming RS. Much of the AL literature suggests that this early AL performance zone is precisely to be expected ([9]), and thus this comparison may often be partially flawed. Further, this same case shows that the AUA (see [5]) would overstate the AL performance gain; see Figure 2a which shows the score trajectories.

Here we resolve that problem by considering the score differences $\delta\mathbf{S}_i$, not the scores themselves \mathbf{S}_i . Those differences show much less autocorrelation than the scores (this is shown by ACF graphs).

An example of the score differences $\delta\mathbf{S}_i$ is shown in Figure 2b. Our new methodology is based on examining these score differences.

A new methodology to evaluate AL performance Our new methodology is based on comparing the score differences δS_i , as a way to compare AL against its benchmark RS. This is done in two stages.

The first stage is to seek a function that quantifies the result of the comparison between two score differences, δS_i^{SE} for AL method SE and δS_i^{RS} for RS. To ensure fair comparisons, ties need to be scored differently to both wins and losses. The approach adopted here is to use a simple comparison function f :

$$f(x, y) = \begin{cases} 1 & : x > y \\ 0.5 & : x = y \\ 0 & : x < y. \end{cases}$$

This comparison function f is applied to two score differences, e.g. $f(\delta S_i^{SE}, \delta S_i^{RS})$. The motivation here is to carefully distinguish wins, losses and ties, and to capture those three outcomes in one scalar summary. Applying that comparison function to compare all the score differences of SE and RS generates a set of comparison values, denoted C_i , each value $\in [0, 1]$. Several instances of RS generate several such sets of values, one for each instance.

We use several instances of RS to capture its high variability, the number of RS instances being N_{RS} . Each instance j has its own set of comparison scores C_i^j . Those comparison values C_i^j are then averaged to form a single set of averaged comparison values, denoted $A_i = \frac{1}{N_{RS}} \sum_{j=1}^{N_{RS}} C_i^j$. Further, each value $A_i \in [0, 1]$.

That single set of values A_i provides a summary of the overall performance comparison between the AL method and RS. That comparison is illustrated in Figure 3 which shows those average comparison values A_i over the whole budget range.

The final stage of the new method is interpreting the averaged comparison values A_i . The aim is to extract the relationship between A_i and budget, with a confidence interval band.

The lower 80% confidence interval is chosen to form a mildly pessimistic estimate of AL performance gain. We fitted a Generalised Additive Model (GAM) to this set of values (given the need for inference of confidence intervals). The GAM is chosen using a logit link function, with variable dispersion to get better confidence intervals under potential model mis-specification (see [6]). The GAM is implemented by R package mgcv version 1.7-22; the smoother function default is thin plate regression splines. The GAM relates the expected value of the distribution to the covariates thus:

$$g(E(Y)) = \beta_0 + \beta_1 f_1(x_1).$$

The fitted GAM is shown in Figure 3. The estimated effect seems roughly linear. The baseline level of 0.5 is shown as a dotted line, which represents an AL method that ties with RS, i.e. does not outperform it.

Given the intricacies of evaluating AL performance, a primary goal for this methodology is to quantify AL performance from a given experiment as a single result. The GAM curve shows where the AL performance zone is, namely the

initial region where the curve is significantly above 0.5. We consider the initial region, as much AL literature suggests that the AL performance gain occurs early in the learning curve, see [9]. Thus the length of the AL performance zone is the single result that summarises each experiment.

Overall, this methodology addresses some of the complexities of assessment of AL performance in simulation contexts. As such it provides a milestone on the road to more accurate and statistically significant measurements of AL performance. This is important given that many authors find that the AL performance effect can be elusive (e.g. [5,3]).

This methodology is illustrated with specific results in Figures 2a, 2b and 3.

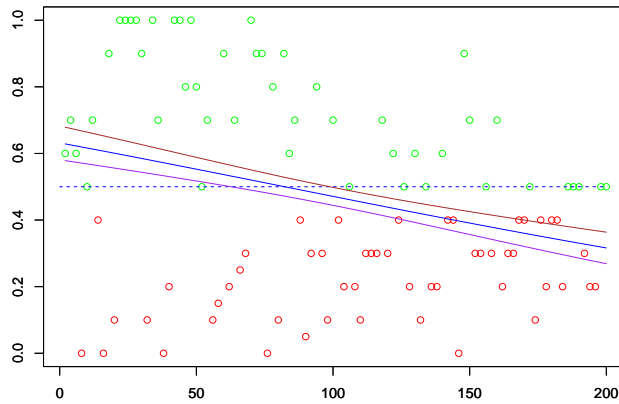


Fig. 3: Averaged Comparison Values A_i with Generalised Additive Model curve and pointwise 80% confidence interval

4 Results and Discussion

The dependent variable is the AL performance zone length, an integer count. That value is obtained via the methodology described above, which includes fitting a GAM to ensure statistical significance. The factors are given in Table 1.

4.1 Negative Binomial Regression Analysis

The experimental output includes the AL performance zone length (derived from the GAM) to the AL factors. Given the form of the aggregate experimental

output, the appropriate initial analyses were Poisson and Negative-Binomial regression. A Poisson regression model was found to be over-dispersed. We fit a negative binomial regression generalised linear model, which fits reasonably well with significant under-dispersion:

$$Y_i \sim \text{NegBin}(\mu_i, \kappa)$$

with

$$\log(\mu_i) = \mathbf{x}_i \cdot \boldsymbol{\beta}$$

where κ is a dispersion parameter.

The significant results of that model are summarised in Table 2.

Table 2: Negative Binomial Significant Results

Name	Coefficient	p-value
Intercept	-1.695	1.70e-12
Specific Classifier, Logistic Regression	1.142	<2e-16
Specific Task, labelled sd7	-0.481	0.000405
Input Type, Continuous	0.578	1.11e-09
Input Type, Discrete	-1.235	<2e-16

There are several results from the negative binomial regression which were not anticipated. For example, LogReg shows more improvement, all things being equal, than SVM, for AL method SE.

This may be due to classifier mis-match: one might conjecture that AL works better when the classifier is mis-matched to the task, because the range of example quality within the unlabelled pool might be much higher under mis-match.

Here classifier mis-match means the experimental metric of the classifier’s sub-optimality on a given task. Classifier mis-match is the performance difference between this classifier and the optimal Bayes classifier on the task. Informally, mis-match measures how ill-suited a classifier is to a given task.

Under correct classifier match, most examples will improve a classifier’s performance, whereas under mis-match, some examples may reduce the performance while others improve it, leading to a greater range of example quality under mis-match.

The choice of task is significant: the third task is worse than the fourth. The fourth task has a more complex decision boundary than the third, leading to expected greater model mis-match for this task. The fact that the third is worse for AL than the fourth is also consistent with the conjecture described above, that AL works better under mis-match.

There is a widespread belief in the AL literature that the AL performance zone is early in the budget range (see [9]). In other words, as we progressively increase the amount of the labelled data, AL provides its performance gain earlier more than later. AL methods are expected to select the more useful examples

from the pool, and the greatest range of usefulness would exist early on. In practical applications, AL is usually required work earlier rather than later, since the essential context of AL is label scarcity. This belief is confirmed by the analysis: for the experiments that showed an AL performance gain, the mean and median lengths of the AL performance zone length were 38 and 32 respectively, out of a maximum of 200.

It is notable that input dimension turns out not to be significant.

It was quite rare for AL to show a performance gain at all, compared to RS, only in around 11% of experiments. This confirms existing studies that the AL performance gain is often elusive ([2,3,8]). It also emphasises the clear need for a precise reasoned methodology to analyse AL performance, hence the detailed methodology described in Section 3.1.

4.2 Results from QBC

To explore the importance of the AL method used, experiments evaluated a different AL method, QBC, using average KL-divergence as the disagreement measure. The two AL methods SE and QBC are very different in both algorithmic details and overall motivation (see [9,4]), making it worthwhile to compare their results.

The AL method QBC takes a committee of classifiers, and scores an unlabelled example \mathbf{x}_j by how much disagreement there is within the committee. Disagreement measures include Vote Entropy and Average K-L Divergence; see [4,9]. For QBC the classifier committee was Logistic Regression, k -nearest-neighbour (with $k = 5$ and $k = 21$), Support Vector Machine and Random Forest.

The experimental setup was identical, and the results were analysed in the same way: by a negative binomial regression analysis. That model fits reasonably well. The results from QBC are somewhat different to those from the SE.

The QBC analysis confirms that input type is significant, with continuous input giving significantly greater AL performance than mixed; and mixed significantly greater than discrete. This confirms that a discretised task is harder than a continuous one, with discretisation reducing the the diversity of pool examples.

It is interesting that two very different AL methods lead to similar results for how AL performance depends on specific factors. We may explain this behaviour in part as follows. With Active Learning there are two distinct stages: firstly the selection of examples for labelling, and secondly the use of those examples in training a particular classifier. With SE the same classifier is used for both stages, whereas QBC uses a classifier committee for selection. The QBC results found that classifier was not significant, in contrast to the SE results which found that Logistic Regression is significantly better than SVM.

This suggests that QBC may be selecting examples which are useful independently of the classifier: good datapoints which benefit any classifier. That in itself is interesting, as it is a very plausible prior belief that the quality of datapoints would be strongly classifier dependent.

5 Conclusion

There are two central questions: Where does AL work? How much does it help? By examining a variety of experiments across a range of points in AL factor space, some conclusions can be drawn.

Overall AL failed to demonstrate a performance gain far more often than not (11% for SE, 6% for QBC). This is consistent with several other authors who reported largely negative results using AL ([1,8]). The analysis also confirmed the general belief in the literature that AL provides its performance gain early on in the budget range. Both AL methods, SE and QBC, showed that the smoothness of the input type makes a significant difference to AL performance.

In future we will extend this work, for example by including many more datasets, some from simulated data, other from real applications, e.g. [5]. Future results should enable recommendations of AL method for applications, by relating the type of classification task to the relative performances of different AL methods.

This experimental study has generated some unexpected results about the factors that determine where AL works. This study has shown many complexities with the assessment of AL performance. It has contributed a new methodology to assess AL performance.

5.1 Acknowledgement

The work of Lewis P. G Evans is supported by an EPSRC doctoral training award.

References

1. Francis Bach. Active learning for misspecified generalized linear models. Technical report, Centre de Morphologie Mathématique, 2006.
2. Jason Baldridge and Miles Osborne. Active learning and the total cost of annotation. *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 1:9–16, 2004.
3. Gavin Cawley. Baseline methods for active learning. *Journal of Machine Learning Research Workshop and Conference Proceedings*, 15:47–57, 2011.
4. Yifan Fu, Xingquan Zhu, and Bin Li. A survey on instance selection for active learning. *Knowledge & Information Systems*, 35:249–283, 2013.
5. Isabelle Guyon, Gavin Cawley, Gideon Dror, and Vincent Lemaire. Results of the active learning challenge. *Journal of Machine Learning Research*, 16:19–45, 2011.
6. Trevor Hastie and Robert Tibshirani. *Generalized Additive Models*. Chapman & Hall / CRC, 1990.
7. Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2nd edition, 2009.
8. Foster Provost and Josh Attenberg. Inactive learning difficulties employing active learning in practice. *ACM SIGKDD*, 12:36–41, 2010.
9. Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.